In [1]:
```
import IPython.display
IPython.display.display_latex(IPython.display.Latex(filename="../macros.tex"))
```

Кластеризация (Clustering)

Обучение без учителя Unsupervised learning

Цель Разбить выборку на непересекающиеся подмножества, которые называются кластеры. Объекты из одного кластера должны быть "похожи", а из разных кластеров "не похожи".

Зачем:

- сжатие данных, оставляем по одному типичному представителю кластера
- обнаружение нетипичности или новизны (novelty detection), обнаружение выбросов.
- понимание данных, упростить дальнейшую работу, работая с каждым кластером отдельно.
- ...

Где исрользуется:

- Marketing
- Biology (упрощение закономерностей взаимодействия генов по репликации ...)
- Informatics (поисковые запросы, сегментирование картинок ...)
- Социология
- ...

мера "похожести" $\rightarrow$ **расстояние** $\rightarrow$ **метрика**

- Euclidean distance (часто используемая)
$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$
- square of the Euclidean (больший вес более отдаленным друг от друга объектам)
$$\sum_{i=1}^{n}(x_i - y_i)$$
- "Manhattan" distance или расстояние городских кварталов (влияние отдельных больших разностей\выбросов уменьшается)
$$\sum_{i=1}^{n}|x_i - y_i|$$
- "Chebyshev" distance (по одной координате)
$$max(|x_i - y_i|)$$
- степенное расстояние
$$\sqrt[r]{\sum_{i=1}^{n}(x_i - y_i)^p}$$

Выбор метрики очень важен, результаты кластеризации могут сильно отличаться.

Схема решения задачи на кластеризацию

Выделение характеристик $\rightarrow$ выбор метрики $\rightarrow$ выбор алгоритма кластеризации $\rightarrow$ разбиение объектов выборки на группы

Формальная постановка задачи кластеризации.

X - множество объектов
Y - множество номеров (меток) кластеров
$\rho(x_1, x_2)$ - функция расстояния между объектами
X^l - конечная обучающая выборка

Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. Каждому объекту приписывается номер кластера.

Решение задачи кластеризации неоднозначно:

- Не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию «по построению». Все они могут давать разные результаты.
- Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием.
- Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

Эвристические алгоритмы

Графические алгоритмы. вершины - объекты, ребра - расстояния.

Алгоритм кратчайшего незамкнутого пути

- Найти пару точек (i, j) с наименьшим $\rho(i, j)$ и соединить их
- пока в выборке остаются изолированные точки
  - найти изолированную точку, ближайшую к некоторой неизолированной
  - соединить эти две точки
- удалить $k - 1$ самых длинных рёбер

Алгоритм FOREL формальный элемент

Идея:

Выбирается точка $x_0 \in X$ и R - параметр. Выделяются все точки внутри сферы с центорм в x_0 и радиуса R. Затем x_0 смещается в центр масс выделенных точек. Повторяем пока не зафиксируется x_0, при этом сфера перемещается в место локального сгущения точек. Все точки в сфере пометим как кластеризованные и удалим из выборки, повторирм пока не пометим все точки.

Потом применяем КНП к центрам. Каждый объект приписываем к кластеру с ближайшим центром.

функционалы качества

Можно рассматривать задачу, как задачу минимизации

within-sum-of-squares

$$wss = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |x_{ij} - c_i|^2$$

k - nb of clusters
n_i - nb of objects in i-th clusters
c_i - center i-th cluster
x_{ij} - j-th object of i-th cluster

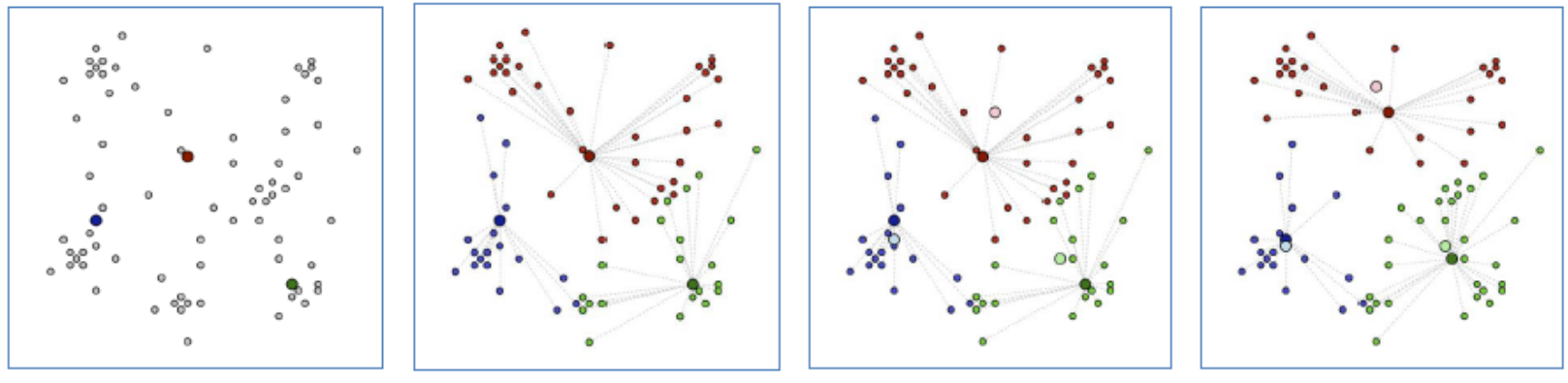
k-means (к средних)

самый распрастранненый алгоритм

1. fix n - number of clusters
2. choose (random) center of clusters
3. clustering
4. repeat:
    - compute center of clusters
    - clustering

Критерий остановки:

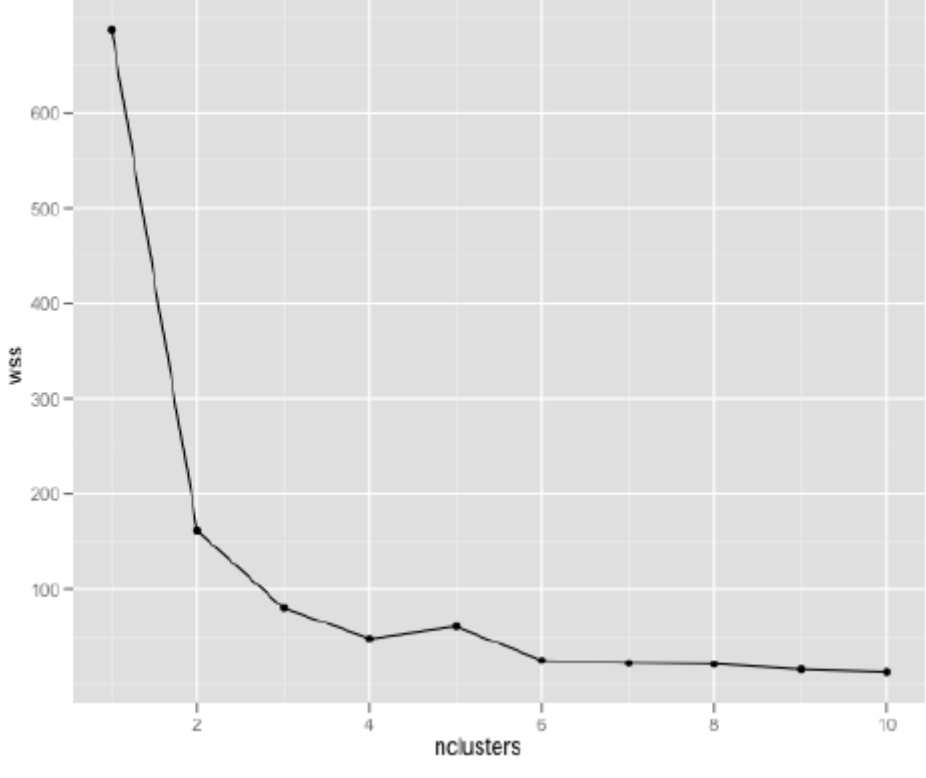
- нет перехода объектов из кластера в кластер
- $diff(wss) < \epsilon$

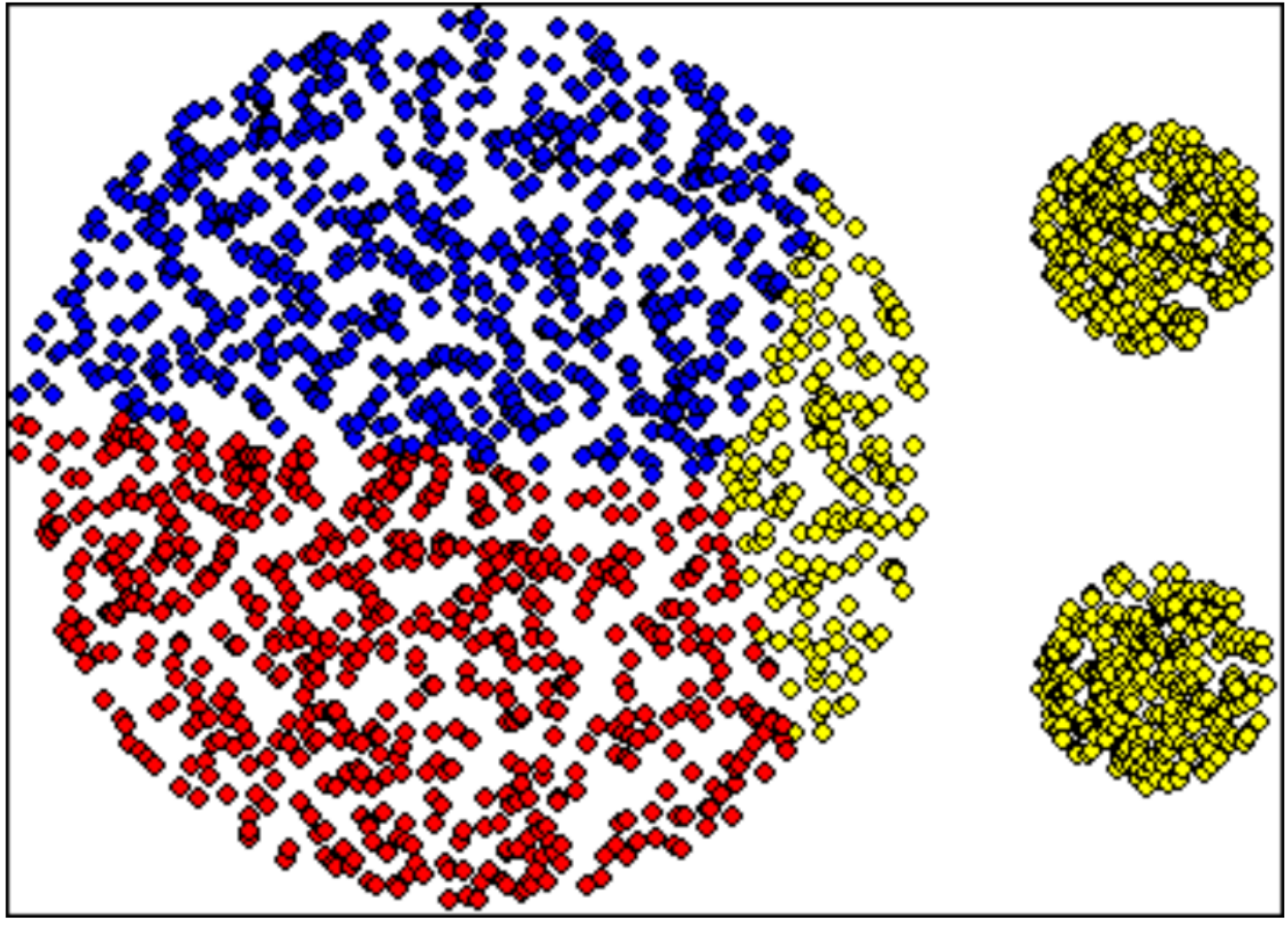


number of clusters

$$wss = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |x_{ij} - c_i|^2$$

Смотрим на уменьшение wss.





k-means

- легко реализовать
- легко кластеризовать новый объект
- на выходе есть "центры" кластеров
- чувствителен к начальному выбору центров масс (запускать несколько раз)
- нельзя использовать категориальные фичи (алгоритму нужна метрика)
- шкалы должны быть одинаковыми (нормализация)
- необходимо подбирать количество кластеров
- "круглые" кластеры

Fuzzy C-Means Clustering

c - number of clusters.
$m \in [1, \inf)$ - exponential weight (1 - k-means, $\inf$ fully fuzzy, all objects to all classes equally)

u_{ij} - matrix. Degree of j-th object to be in i-th class.

$$u_{ij} \in [0, 1], i \in [1, c], j \in [1, N]$$

$$\sum_{i=1}^{c} u_{ij} = 1, \ for\ all\ j$$

- each object in all clusters

$$0 < \sum_{j=1}^{N} u_{ij} < 1$$

- each cluster not empty and does not contain all elements

$$J = \sum_{i=1}^{c} \sum_{j=1}^{N} (u_{ij})^m \rho(v_i, x_j)$$

x_j - j-th object in cluster
$v_i = (v_{i1}, \ldots, v_{iM})$ - center of i-th clusters:

$$minimize(J)$$

$$u_{ij} = \begin{cases} 1, if\ i = j\ and\ d_{ij} = 0 \\ 0, if\ i \neq j\ and\ d_{ij} = 0 \\ \dfrac{1}{d_{ij}^{\frac{2}{m-1}}\ \sum_{k=1}^{c} \frac{1}{d_{kj}^{\frac{2}{m-1}}}}, if\ d_{ij} > 0 \end{cases}$$

$$d_{ij} = \rho(v_i, x_j),\ i \in [1, c],\ j \in [1, M]$$

Algorithm

- fix c, m, ϵ - stop param
- Initialize $U^{(0)} = [u_{ij}]$ matrix
- repeat until $||U^{(n-1)} - U^{(n)}|| > \epsilon$:
  - compute centers $v_i^{(n)}$
  - compute $U^{(n)}$ by centers

Hierarchical clustering

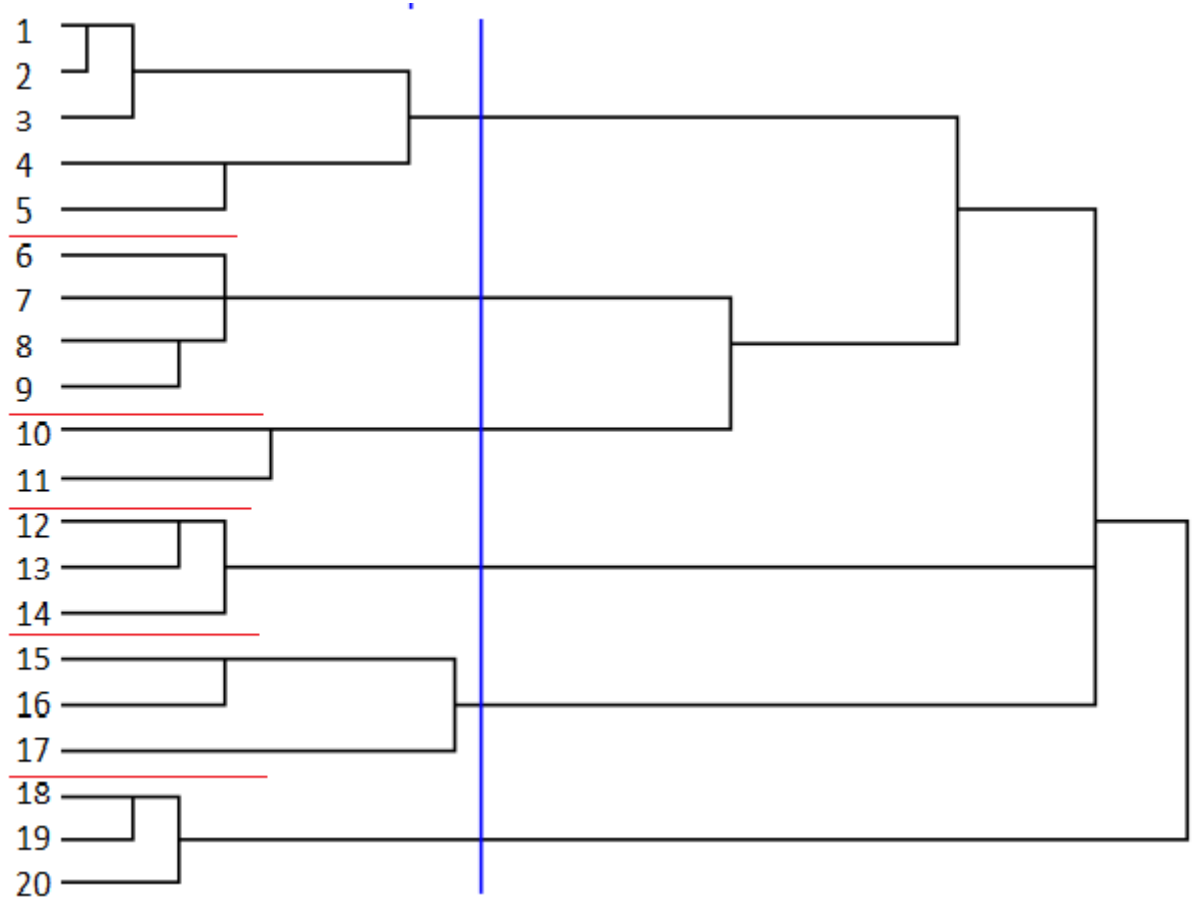
Таксономия.

Заключается в построении дендрограммы.

How to build a dendrogram?(снизу вверх)

At each step, we find the two closest clusters and merge them until all the elements are collapsed into one cluster.

You can cluster both the number of clusters and the distance between clusters.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

расстояние между кластерами

- расстояние между ближайшими элементами. Тенденция объединяться в цепочки.
- Расстояние между далекими элементами. Этот метод обычно работает очень хорошо, когда объекты происходят из отдельных групп. Если же кластеры имеют удлиненную форму или их естественный тип является «цепочечным», то этот метод непригоден.
- Расстояние между центрами кластеров.
- Невзвешенное попарное среднее. В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты формируют различные группы, однако он работает одинаково хорошо и в случаях протяженных («цепочечного» типа) кластеров.

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)

$\rho(x, y)$ - distance function
fix ϵ and m

define:
$E(x) = \{y \mid \rho(x, y) \leq \epsilon\}$ - окркстность объекта.
Корневым объектом - such x that $|E(x)| \geqslant m$
p - **непосредственно плотно-достижим** из q если $p \in E(q)$ и q - корневой объект
p - **плотно-достидим** из q если
$\exists p_1, p_2, \ldots, p_n, p_1 = q, p_n = p \; and \; \forall i \in [1, .. n] \; : \; p_{i+1}$ непосредственно плотно-достижимы из p_i

Now if p is a core point, then it forms a cluster together with all points (core or non-core) that are reachable from it.

DBSCAN

- не нуждается в количестве кластеров
- может находить кластеры различной формы
- устойчиво к выбросам
- зависит от расстояния
- нужны 2 параметра ϵ и m